# Bi-modality Fusion for Emotion Recognition in the Wild

**Sunan Li**
School of Information Science and Engineering,
Southeast University
Nanjing, China
230189473@seu.edu.cn

**Wenming Zheng**[*]
Key Laboratory of Child Development and Learning Science of Ministry of Education
School of Biological Science and Medical Engineering,
Southeast University
Southeast University, China
wenming_zheng@seu.edu.cn

**Yuan Zong**
School of Biological Science and Medical Engineering,
Southeast University
Southeast University, China
xhzongyuan@seu.edu.cn

**Cheng Lu**
School of Information Science and Engineering,
Southeast University
Southeast University, China
cheng.lu@seu.edu.cn

**Chuangao Tang**
School of Biological Science and Medical Engineering,
Southeast University
Southeast University, China
tcg2016@seu.edu.cn

**Xingxun Jiang**
School of Biological Science and Medical Engineering,
Southeast University
Southeast University, China
jiangxingxun@seu.edu.cn

**Jiateng Liu**
School of Biological Science and Medical Engineering,
Southeast University
Southeast University, China
jiangxingxun@seu.edu.cn

**Wanchuang Xia**
School of Cyber Science and Engineering,
Southeast University
Southeast University, China
220184474@seu.edu.cn

## ABSTRACT

The emotion recognition in the wild has been a hot research topic in the field of affective computing. Though some progresses have been achieved, the emotion recognition in the wild is still an unsolved problem due to the challenge of head movement, face deformation, illumination variation etc. To deal with these unconstrained challenges, we propose a bi-modality fusion method for video based emotion recognition in the wild. The proposed framework takes advantages of the visual information from facial expression sequences and the speech information from audio. The state-of-the-art CNN based object recognition models are employed to facilitate the facial expression recognition performance. A bi-direction long short term Memory (Bi-LSTM) is employed to capture dynamic information of the learned features. Additionally, to take full advantages of the facial expression information, the VGG16 network is trained on AffectNet dataset to learn a specialized facial expression recognition model. On the other hand, the audio based features, like low level descriptor (LLD) and deep features obtained by spectrogram image, are also developed to improve the emotion recognition performance. The best experimental result shows that the overall accuracy of our algorithm on the Test dataset of the EmotiW challenge is 62.78%, which outperforms the best result of EmotiW2018 and ranks 2nd at the EmotiW2019 challenge.

## CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

## KEYWORDS

Emotion Recognition ; Deep Learning ; Convolutional Neural Networks

[*] The corresponding author.

*ICMI'19, October 14-18, 2019, Suzhou, China*

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3355719

**ACM Reference Format:**
Sunan Li, Wenming Zheng, Yuan Zong, Cheng Lu, Chuangao Tang, Xingxun Jiang, Jiateng Liu, and Wanchuang Xia. 2019. Bi-modality Fusion for Emotion Recognition in the Wild. In *2019 International Conference on Multimodal Interaction (ICMI'19), October 14-18, 2019, Suzhou, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3340555.3355719

## 1 INTRODUCTION

Emotion recognition plays a key role in human-computer interaction and has been investigated for many years. There are several types of emotional data modality, including facial expression, audio emotion, EEG, EMG and etc. Among these data modalities, a recorded video clip usually contains both human facial expression and audio. Video data contains more emotional information compared with a static facial image. Besides, video recording has an advantage of non-contact which leads to its widespread application. With the development of social media and internet, massive amount of videos can be used. As a result, audio-video based emotion recognition has attracted many researchers' interests.

The audio-video based emotion recognition sub-challenge in the wild (EmotiW) challenge has been held for seven times since 2013, which provides a benchmark for researchers. The video clips used in this challenge were extracted from movies to simulate emotions in real world environment [1, 2]. And many researchs have been done in EmotiW challeng[4, 18]. However, the video-based real world facial expression recognition suffers from the challenge of illumination variations, face occlusion etc. To train a robust emotion recognition model, researchers have proposed several kinds of methods, including hand-crafted features and deep learning based features. Kaya et al. [8] employed traditional features and least square based learners for emotion recognition. Wu et al. [19] used a bunch of features model to encode video clips. There are also some literatures focusing on hand-crafted features (e.g. Gabor filters, Local Binary Patterns) [10, 21], which are then fed into classifiers such as SVM or Random Forest.

In the past years, researchers have been working on convolution neural network (CNN) for the tasks of computer vision and pattern recognition. A great success has been achieved in these fields. In the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), Alexnet achieved a best overall accuracy of 84.7%[9]. Motivated by the Alexnet, GoogLeNet[17], VGG[16] and ResNet[5] have been proposed to tackle problems in visual tasks. To capture dynamic information of the facial expression in the sequences, recurrent neural networks has been developed to tackle this problem. In the 2018 EmotiW challenge, Liu et al. [11] employed a multi-feature based framework using acoustic features and facial features in both temporal mode and non-temporal mode. The framework achieved an accuracy of 61.87% on the test set and ranked first in the last audio-video emotion recognition sub-challenge. Motivated by their work, in this paper we propose a framework containing facial image model and audio model. The image model is consisted of four different types of neural networks to extract features and then the features are fed into Bi-direction LSTM to capture dynamical temporal information for a higher accuracy [3]. In addition, we extract the complementary information from the audio with two different methods. In the subsequent fusion stage, the grid-search strategy is employed for performance optimization on the validation set, and the information fusion method improve the accuracy to 62.78% on the test set and the performance ranks 2nd in the audio-video based sub-challenge in the EmotiW2019.

The remainder of this paper is organized as follows. The model of facial image and audio is presented in section 2. In section 3, we expound our experimental results on the challenge dataset to evaluate our proposed framework. Finally in section4, we conclude this paper.

## 2 THE PROPOSED METHOD

Our proposed framework consists of two parts, i.e. image flow and audio flow. As for image flow, there are three CNN-based networks for unique spatio-temporal feature learning, while the rest one focuses on spatial information extraction. Both the scores from image flow and audio flow are weighted with a grid-search optimization. The final weighted score is used for classification. The framework is shown in Fig. 1 and the details are described as follows.

### The Facial Image Model

The cropped facial sequences are fed into the convolutional neural networks, i.e. VGG-Face, Restnet18, Densenet121 and VGG16, for spatial feature extraction. In the first three networks, the stacked convolutional neural networks are connected with a two-layer Bi-LSTM[6, 12], which can capture dynamic information. As for the static CNN model, the predicted label of each frame is used for calculation of frequency and the normalized frequency is used as a score of each emotion type.

*CNN-Based Spatial Feature Extraction Network.* In ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014, Visual Geometry Group proposed the VGGNet and ranked 2nd in classification sub-challenge. VGG-Face model is a special type of VGG with 16 layers that trained on a special database which is developed by visual geometry group in oxford university too and trained on labeled faces in the wild and the youtube faces [14]. It has achieved a great success in the field of face recognition in recent years. Therefore we use the VGG-Face here to capture the emotion features of facial images. In order to obtain better recognition result, we

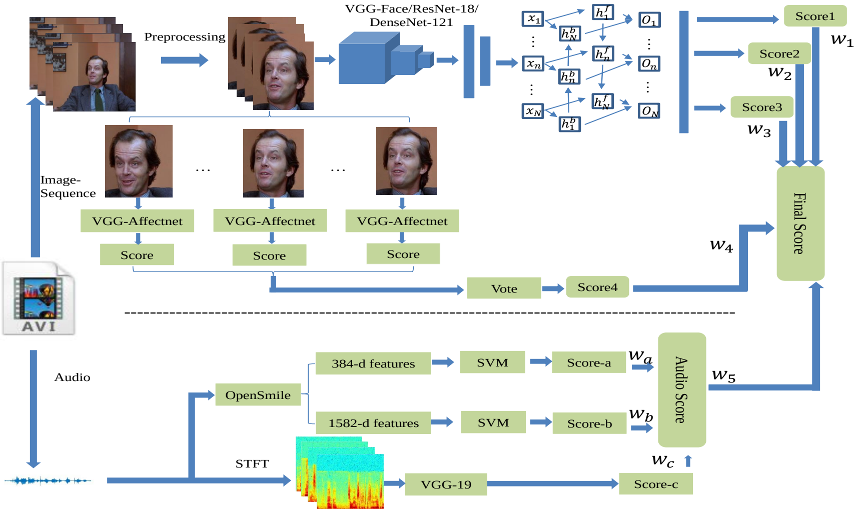

**Figure 1: The overview of proposed framework.**

fine tune the pre-trained model in facial expression images cropped from training dataset. As previous research has been done by Yosinski et al in Cornell University, the features extracted from deeper layers of networks tend to specific on tasks than shallow layers[22]. Therefore, we freeze the weights of all convolutional layers and only update that in the fully connected layers in VGG-Face network by the processing facial images.

While with the network depth increasing, two problems are unavoidable. The gradients may be vanish/explode, and the accuracy may be saturated and then degrade rapidly. To address these two problems, He et al. proposed the deep residual framework called 'ResNet'. ResNet achieved the first place in the annual competition ILSVRC-2015[5]. The 'shortcut connections' conception was first proposed in ResNet and it can change traditional fitting rule. The main role of 'shortcut connections' was to make the input maps be identity mapping. Additionally, we use the ResNet-18 rather than ResNet-50 to construct our framework due to the lack of train data to avoid over-fitting.

The number of parameters of ResNet is substantially larger because each layer has its own weights and the ResNet only use output feature of adjacent bottom layer as input which limit the reuse of feature maps. To solve this problem Huang et al proposed DenseNet which won the best paper award in CVPR-2017[7]. The name DenseNet means each layer obtains additional inputs from all preceding layers and passes on its own feature maps to all subsequent layers, which means they are more densely connected than other networks. Compared with ResNet, DenseNet encourages feature reuse and alleviates the vanishing gradient problem.

The three network mentioned above improve the performance of proposed framework based on network structure. To tackle the problem of sample imbalance , we introduce a new emotion database which called AffectNet. Affect from the InterNet is the largest database of the categorical and dimensional models of affect in the wild[13]. We use it to train a VGG-16 model to improve the robustness of our framework. And the trained VGG-16 model output a score vector correspond to the probability of the input image belong to each emotion. Finally, for the selected frames from one video clip, using simple majority vote to determine the emotion video belongs to.

*Dynamic Feature Extraction Network.* The four networks mentioned above extract the emotion feature both in spatial dimension, now we need a framework to combine features extracted from different images that cropped from one video to capture dynamic information. To tackle this problem, we use LSTM to extract dynamic information, LSTM is a variation of original RNN which can memorize the value over any time interval and control the information flowing into the later sequence. But RNN have some limitation in processing various input, so it's hard to be used in the task of recognize emotion in the wild directly. The core of LSTM is the state of cell, in which there are three gates (input gate, forget gate and output gate) utilized in LSTM to update the cell state to solve the problem of long-term memory problem emerged in RNN [6, 12]. The output of these three gate is shown in formula(4), where $f_t$,$i_t$,$o_t$ represent the output of input gate, forget gate and output gate in time t respectively.

$$
\begin{aligned}
f_t &= \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \\
i_t &= \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \\
o_t &= \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)
\end{aligned} \tag{1}
$$

Finally, the hidden state can be formulate as formula(2), where $C_t$ represent the cellular state at time t.

$$
h_t = o_t * \tanh\left(C_t\right) \tag{2}
$$

To extract the dynamical information from video more robust, we use bi-direction LSTM that use the one direction frame sequence and its reverse as input. So the output of Bi-LSTM can be formulate as:

$$
h_t = h_t^f + h_{T-t}^b \tag{3}
$$

These features constituting the sequences are obtained from the last fully connected layer in CNN, and the dimensions of features are all 4096.

### The Audio Model

Audio is a significantly important part in EmotiW challenge, due to the lack of samples of some specific emotions in training database and the similarity of input, different image model perform so similarly that hard to classify some sample and emotion correctly. The learning of audio signals plays an important role in improving the performance of our video emotion classification model.The key step of audio based emotion recognition include pre-processing, feature extraction and classification. In our framework, we remove the background noise and filter the futile part to extract the useful speech from video as the pre-processing. In this framework two methods were used as feature extraction and classification, first one is a deep learning model in which we extract the useful speech from video and convert it into spectrogram by short time fourier transform to extract the information in time and frequency domains which contain rich information of different emotions. Then processing the spectrogram as a image and fed it into a VGG-19 network[15] to extract depth information and classified by softmax. Secondly, to improve the robustness of audio model, we also use the openSMILE toolbox to extract LLD feature that is commonly utilized in audio processing. We extract two different kind of openSMILE features whose dimensions are 384 and 1582 respectively. These two features are two standard feature contain Mel Frequency Ceptral Coefficien (MFCC), fundamental frequency etc.. Then using Support Vector Machine (SVM) to processing these two features and output two corresponding scores.

**Table 1: Recognition accuracy of each model on the validation and test database.**

| Models | Validation(%) | Test(%) |
|---|---|---|
| VGG-Face+BLSTM | 53.91% | - |
| ResNet-18+BLSTM | 43.34% | - |
| DenseNet-121+BLSTM | 49.35% | - |
| VGG-AffectNet | 41.78% | - |
| Audio-Fusion | 25.59% | - |
| Fusion | 54.30% | 62.78% |

Finally, we fusion three different score that come from spectrogram image and openSMILE respectively as final speech score, the fusion strategy will performed later. Therefore we not only take the strong advantage of deep feature based on time-frequency representation of spectrogram and the capacity on image processing of CNN, also using the LLD features classified by SVM to improve the robustness of our audio model.

### The Fusion Strategy

In previous sections, we present different network we used in our framework, each of them play an important and complementary role in classify the emotion. So we have to use a strategy to fusion the output of each networks. In our framework, each network will output a score vector for each sample which donating the probability of the sample belongs to the specific emotion. In this paper we use weighted sum to fusion every score and get the final score as shown in formula(4).

$$
\begin{aligned}
\hat{S} = \omega_1 \cdot S^{VGG-FACE} &+ \omega_2 \cdot S^{\mathrm{ResNet}} + \omega_3 \cdot S^{DenseNet} \\
&+ \omega_4 \cdot S^{VGG-AffectNet} + \omega_5 \cdot S^{Audio}
\end{aligned} \tag{4}
$$

Where $\hat{S}$ is the final score vector and $\omega_i$ and $S$ are the weight and score vector of network i respectively. For better performance, we use grid search method on validation database to find the proper weights, and the final weight is 0.88,0.34,0.22,0.06,0.3 for VGG-Face, ResNet-18, DenseNet-121, VGG-Affectnet and audio respectively.

## 3 EXPERIMENT

### Pre-Processing

The dataset of audio-video sub-challenge of EmotiW2019 is collected from movies and TV series, which similar to the emotion in the wild, many of the video clips are affected by abnormal conditions such as illumination variation, face occlusion and so on. Therefore, pre-processing is needed to process the video to generate frames we used in our framework. So we alignment and normalization the face through the identification block of the Seetaface using five landmarks getting in MTCNN and resize the facial image into 256×256[20, 23]. Then we crop the image into 224×224 randomly. Finally, for each video, we select 16 frames as input for the ResNet18 VGG-Face and VGG-AffectNet, 4 frames for DenseNet-121.

Similarly, for the audio, we first removing the irrelevant and background noise to get the speech. Then the speech is transformed into spectrogram images by short time fourier transform and resized to 224×224 for the VGG-19 modal.

### Result and Discussion

In the audio-video sub-challenge in EmotiW2019, we adopt different neural networks to extract complementary features for the better and robust performance. The results of different network we proposed in section2 was shown in Table1 respectively, which the best model: VGG-Face have achieve overall accuracy of 53.91%. It's worth noting that though the accuracy of audio-fusion model lower than other networks, the audio-fusion model still plays an important role due to its complementary information. Finally, the best fusion framework achieve 62.78% on test database, 0.91% higher than the champion of EmotiW2018.

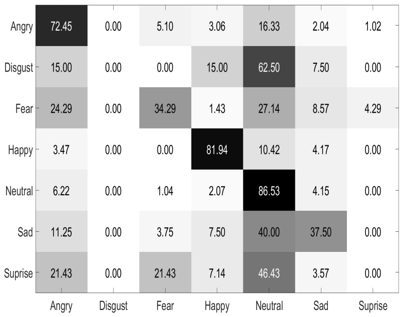

**Figure 2: Confusion matrix of the 4th submission.**

The confusion matrix of fusion networks is shown in Fig. 2. According to the confusion matrix shown, there are three emotions (disgust, fear and surprise)are hard to classify correctly, perhaps due to the lack of training data and the implicity of features of these three emotion. All of the five submission have better performance on test database than validation database, perhaps due to the Proportion of hard classified emotion is lower in test database than in validation database.

## 4 CONCLUSION

In this paper, we present a bi-modality multi feature framework to recognize emotion in the wild in EmotiW2019. Emotions information are split into two complementary aspect: audio and video. For video information we use four different network to extract emotion feature, And for audio, we using STFT and openSMILE separately. Then we combine different scores by weighted sum where the weight of each network depended on their contribution to the best result. The experiment result of the challenge showed that out proposed framework is more robust on the task of emotion recognition in the wild.

## 5 ACKNOWLEDGE

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, in part by the National Basic Research Program of China under Grant 2015CB351704, in part by the National Natural Science Foundation of China under Grant 61572009, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242018K3DN01 and Grant 2242019K40047.

## REFERENCES

[1] Abhinav Dhall, Roland Goecke, Shreya Ghosh, and year=2019 publisher=ACM Tom Gedeon, booktitle=ACM International Conference on Mutimodal Interaction 2019. [n.d.]. EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. ([n. d.]).

[2] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia* 19, 3 (July 2012), 34–41.

[3] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005), 602 – 610. IJCNN 2005.

[4] Da Guo, Kai Wang, Jianfei Yang, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. 2019. Exploring Regularizations with Face, Body and Image Cues for Group Cohesion Prediction. In *Proceedings of the 21th ACM International Conference on Multimodal Interaction (in press)*. ACM.

[5] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[7] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[8] Heysem Kaya, Furkan Gürpinar, Sadaf Afshar, and Albert Ali Salah. 2015. Contrasting and Combining Least Squares Based Learners for Emotion Recognition in the Wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, 459–466.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60, 6 (May 2017), 84–90.

[10] Markus Kächele, Martin Schels, Sascha Meudt, Günther Palm, and Friedhelm Schwenker. 2016. Revisiting the EmotiW challenge: how wild is it really? *Journal on Multimodal User Interfaces* 10, 2 (2016), 1–12.

[11] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. 2018. Multi-Feature Based Emotion Recognition for Video Clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*. ACM, New York, NY, USA, 630–634.

[12] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2015. Recurrent neural network based language model. In *Interspeech, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September.*

[13] A. Mollahosseini, B. Hasani, and M. H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in

the Wild. *IEEE Transactions on Affective Computing* 10, 1 (Jan 2019), 18–31.

[14] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference*.

[15] R V Shannon, F G Zeng, . Kamath, V., . Wygonski, J., and . Ekelid, M. 1995. Speech recognition with primarily temporal cues. *Science* 270, 5234 (1995), 303–304.

[16] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science* (2014).

[17] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9.

[18] Kai Wang, Jianfei Yang, Da Guo, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. 2019. Bootstrap Model Ensemble and Rank Loss for Engagement Intensity Regression. In *Proceedings of the 21th ACM International Conference on Multimodal Interaction (in press)*. ACM.

[19] Jianlong Wu, Zhouchen Lin, and Hongbin Zha. 2015. Multiple Models Fusion for Emotion Recognition in the Wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, 475–481.

[20] Shuzhe Wu, Meina Kan, Zhenliang He, Shiguang Shan, and Xilin Chen. 2017. Funnel-structured cascade for multi-view face detection with alignment-awareness. *Neurocomputing* 221 (2017), 138 – 145.

[21] Anbang Yao, Junchao Shao, Ningning Ma, and Yurong Chen. 2015. Capturing AU-aware facial features and their latent relations for emotion recognition in the wild. In *Acm on International Conference on Multimodal Interaction*.

[22] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *Eprint Arxiv* 27 (2014), 3320–3328.

[23] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (Oct 2016), 1499–1503.